Disaggregating between- and within-classroom variation in student behavior

Disaggregating between and within-classroom variation in student behaviour: A multi-level factor analysis of teacher ratings of student prosociality and aggression

Abstract

Teacher ratings of student behaviours vary systematically both at the student and teacher/classroom level. Multi-level confirmatory factor analysis (ML-CFA) can disaggregate between- and within-teacher/classroom variance, identify an optimal psychometric model at each level and test correlates of the resulting dimensions. In this study, 250 teachers (37% male) rated an average of 4.02 students (51% male; aged 10 at time 1 and 11 at time 2) from a normative sample of Swiss youth. Substantial and unidimensional between-teacher variation in ratings of both prosociality and aggression were identified and this was stable across time. These dimensions were not associated at the between-teacher/classroom level with teacher gender nor teacher-student relationship, although they were associated with teacher relationships at the within-teacher/classroom level. There was little between-teacher/classroom variation observed in student self-reports of prosocial and aggressive behaviour and multi-level CFA was not possible for these ratings. Future research should aim to identify sources of between-teacher/classroom variation. This should include factors that influence negative and positive teacher perceptions of and response biases related to student behaviour as well as student behaviour itself.

Keywords: multi-level confirmatory factor analysis, prosociality, aggression, teacher gender, teacher relationships

It is now widely recognised that classrooms play an important role in shaping children and adolescents not only academically, but also socially and emotionally (e.g. Dunn, Masyn, Jones, Subramanian, & Koenen, 2015). However, there is considerable variation between classrooms in terms of dimensions of youth psychosocial functioning and behaviour (e.g. Fraser et al., 2005; Malti, Ribeaud & Eisner, 2011; O'Brennan, Bradshaw & Furlong, 2014; Obsuth, Sutherland, Cope, Pilbeam, Murray & Eisner, 2016; Salmivalli, Voeten & Poskiparta, 2011). Understanding this variation has thus become an important research goal for informing how teachers and classroom environments can best support youth social and emotional development and prevent behaviours such as aggression and bullying.

In school-based studies of youth psychosocial functioning, the amount of variation at the between-teacher/classroom level varies by study but is often in the range of 5-30%. O'Brennan et al. (2014), for example, reported between-classroom intraclass correlations (ICCs) for teacher-reported problem behaviour of .16. Their sample comprised 467 classrooms in 37 elementary schools across 5 Maryland school districts including 8750 1st to 5th grade students (48% 'White', 44% 'Black', 8% 'Other' and 52% male) from a longitudinal prevention study. Most of the participating teachers were White (89%) and female (92%). Similarly, Salmivalli et al. (2011) reported classroom-level ICCs for self-reported bystander actions in relation to bullying of .21-.35. The data was from the pre-test phase of a Finnish study evaluating the KiVa antibullying programme and comprised 6,764 3rd to 5th grade students (50% male, the vast majority Caucasian) in 385 classes. In the current sample of Swiss youth from the z-proso study (see 'Methods' for details), ICCs for various dimensions of teacher-reported externalising and prosocial behaviour have been reported in past research as averaging .25 (Malti et al., 2011).

Sources of clustering by teachers/classrooms

It is often assumed that clustering of student behaviours by teachers/classrooms reflects the influence of shared classroom environments that induce students to behave more similarly to one another. Classroom environments refer to both context and climate variables. Context refers to a class aggregate of individual student characteristics (e.g. the average of individual aggression or individual

achievement within a classroom). Past research has suggested, for example, that classrooms with higher levels of negative behaviour such as aggression are liable to experience classroom-wide increases in problem behaviour (e.g. Barth, Dunlap, Dane, Lochman & Wells, 2004; Thomas & Bierman, 2006; Thomas, Bierman & Powers, 2011). On the other hand, classrooms with high levels of positive e.g., prosocial behaviour are more likely to show classroom-level increases in prosocial behaviour over time (e.g. Laninga-Wijnen, Harakeh, Dijkstra, Veenstra, & Vollebergh 2017). Climate refers to classroom-level constructs (e.g. classroom organisation; Morin, Marsh, Nagengas, & Scalas, 2012). Arens & Morin (2016), for example, found that levels of teacher emotional exhaustion were associated with between -classroom variation in grades, achievement scores and perceptions of teacher support.

However, teachers often supply ratings of behaviour for multiple students in their class. In these cases, another potential contributing factor to between teacher/classroom variation must be considered. Specifically - like any other informant completing a rating scale - teachers are subject to response biases that affect their ratings over and above the student behaviours that they observe (e.g. Abikoff, Courtney, Pelham & Koplewicz, 1993). Rating tendencies vary between individuals and are affected by characteristics of the rater. For example, higher levels of depression, stress and neuroticism have been shown to be associated with more pessimistic responses to questionnaire items (e.g. Berg-Neilsen & Dahl, 2003; Youngstrom, Loeber & Stouthamer-Loeber, 2000). This means that all else being equal, responses pertaining to the same targets provided by the same rater can resemble one another more than responses provided by different raters. Self-reports by students are also subject to biases but as they typically only report on their own behaviour, individual differences in response style and behaviour are completely confounded and are not expected to contribute to classroom-level ICCs. By extension, ratings of the same behaviour via self- rather than teacher- reports would be expected to yield lower ICCs.

Whether due to classroom context/climate effects or response styles, it is important to properly disaggregate variation in reports of student behaviour that is due to differences between teachers/classrooms (level-2 variation) from variation between students who share a teacher and classroom (level-1 variation). In classroom climate/context research this is necessary to avoid

ecological fallacies by ensuring that individual- and classroom- level effects are not conflated (e.g. see Marsh et al., 2012). Here, proper disaggregation of variance at different levels facilitates the identification of sources of variation between teachers/classrooms which can, in turn, help identify teacher/classroom-level targets for intervention (e.g. Dunn et al., 2015). In psychometric research, disaggregating these two levels of variation could help control for teacher response biases in ratings of student behaviour to yield more accurate individual-level estimates. In addition, teachers with more negative rating tendencies may benefit from intervention to the extent that this reflects the effects of stress/depression and/or negative views of students that are unmerited and may be unhelpful with respect to fostering student academic and psychosocial development.

Multi-level confirmatory factor analysis

Disaggregation of between and within teacher levels of variation can be achieved using multi-level confirmatory factor analysis (ML-CFA) (Muthén, 1994). ML-CFA allows factor structures to be separately tested at between-teacher/classroom and within-teacher/classroom levels when teachers provide ratings on multiple students in their class. It allows for measurement error at the item level as well as sampling error in the aggregation of student-level scores into classroom/teacher level scores to be accounted for.

Identifying an appropriate psychometric model for student prosocial and problem behaviour at the between-teacher/classroom level using ML-CFA provides a foundation for understanding the nature of its between-teacher/classroom variation. The dimensions identified can be tested for their stability over time and their predictors and outcomes to evaluate particular hypotheses regarding teacher/classroom level effects. For example, stable between-teacher variance, especially over different sets of students over multiple years would point to trait-like features of the teacher that affect the classroom environment they contribute to and/or the manner in which they rate students (e.g. a pessimistic bias). Similarly, covarying between-teacher/classroom ratings with other teacher/classroom-level variables such as depression, stress, burnout, training, gender, personality or school climate variables, can help identify specific influences on teacher ratings of student behaviour.

Past research has highlighted the utility of ML-CFA in school research (e.g. Arens & Morin, 2016; Dunn et al., 2015; Downer, Stuhlman, Schweig, Martínez & Ruzek, 2015; Marsh et al., 2012; Morin, Marsh, Nagengast & Scalas, 2014; Spilt, Koomen & Jak, 2012). Dunn et al. (2015), for example, highlighted the importance of testing whether different factor structures are optimal at different levels. Analysing data on 79,362 students in 126 schools from the *Add Health* study, they found that at the within-school level, four latent factors were identified: school adjustment, externalising problems, internalising problems and self-esteem. Variation in the same items at the between-school level was better captured by three dimensions: collective school adjustment, psychosocial environment and collective self-esteem.

ML-CFA is also well-suited to understanding between-teacher/classroom variation in core dimensions of student social behaviour such as prosociality and aggression. Figure 1 depicts a hypothetical ML-CFA model using aggression as an example. It shows five aggression items (rectangles Agg1-Agg5) which serve as indicators of latent between-classroom variables (circles Agg1B-Agg5B) as well as a within-classroom/teacher latent aggression variable (AggW). The Agg1B-Agg5B latent variables are themselves indicators of a latent between-classroom/teacher latent variable (AggB). The single-headed arrows terminating in Agg1-Agg5 and Agg1B-Agg5B represent measurement errors. Here, for simplicity, aggression is represented by a single latent variable both at the within- and between- level. However, it could be better represented using multiple latent variables at either or both levels (e.g. with correlated reactive and proactive latent aggression factors at the within-level and a single general aggression factor at the between level). The optimal factor structure at each level is one question that ML-CFA is well-suited to answer. The figure also includes the kinds of factors that may contribute to variation in teacher reports of student behaviour at the student and classroom level. At the student-level, predictive factors would include student characteristics such as trait self-control, anger, anger rumination, violent ideations, hyperactivity/impulsivity, oppositionality, and conduct problems (e.g. Harty, Miller, Newcorn & Halperin, 2009; Murray, Obsuth, Eisner & Ribeaud, 2016; Murray, Obsuth, Zirk-Sadowski, Ribeaud & Eisner, 2016). At the teacher/classroom level, predictors could include factors affecting teacher response style such as stress, burnout and depression, together

with other teacher traits that impact their ability to minimise disruptive behaviour in the classroom (e.g. Arens & Morin, 2016; Pas & Bradshaw, 2014). In addition, context factors such as the aggregate ability and social norms of the students comprising the class are also potentially important predictors at this level (e.g. Menesini, Palladino & Nocentini, 2015). Many factors may be relevant at both levels, for example, individual and aggregate norms of disruptive behaviour could respectively influence aggression at the individual and classroom-level.

Teacher gender

Some studies have suggested that teacher gender may be a source of between-teacher variation in student behaviour ratings. Hopf and Hatzichristou (1999) found that female teachers tended to give more favourable ratings of 11-year olds' psychosocial and academic adjustment than male teachers and proposed that this could be explained by female teachers being more accepting of misbehaviour. This suggestion was echoed by Spilt et al. (2012) who found that female teachers tend to report better relationships with their primary school students than male teachers. More recently, Pas & Bradshaw (2014) found that male teachers rated emotion dysregulation problems as higher and prosocial behaviour as lower than female teachers.

Teacher-student relationships

Other studies have pointed to teacher-student relationships as potentially relevant to variations in student behaviour at both the between-teacher/classroom and within-classroom level. Like student behaviour, it is useful to conceptualise teacher-student relationships as multi-level construct influenced both by differences between students and by differences between teachers shared by multiple students (e.g. Spilt et al., 2012), with up to a third of the variance in teacher-student relationships deriving from differences between teachers rather than students (Mashburn et al., 2006). In terms of its links with student behaviour, past has suggested that better teacher-student relationships are associated with less problem behaviour and more positive behaviour. For example, using a propensity score matching approach in the current sample, Obsuth, Murray, Malti, Sulger, Ribeaud & Eisner (2017) reported that positive teacher-student relationships promoted less externalising and more prosocial behaviour. Others

have highlighted the potential reciprocal nature of the relationship, identifying cross-lagged associations between externalising behaviour and poorer relationships with teachers (Pakarinen et al., 2017; Theimann, 2016). Teacher relationships and student behaviour are, however, also likely to be related at the between-teacher level. For example, teachers with classes comprising students showing higher levels of aggressive behaviour are more likely to experience greater stress and burnout (e.g. Aloe, Shisler, Norris, Nickerson & Rinker, 2014), which can impact on their ability to form supportive relationships with their students (e.g. Arens & Morin, 2016).

The current study

Given the potential of ML-CFA to illuminate sources of variation between teachers and classroom relevant for student behaviour, we here applied the method to explore the extent, structure, stability, and correlates of between-teacher/classroom variation in ratings of student behaviour. We hypothesised that there would be systematic variation in student behaviour ratings both at the student-level and at the teacher/classroom level when teachers provided ratings of behaviour. We also hypothesised that female teachers would provide more positive ratings than male teachers. We hypothesised that variations across teachers/classrooms in aggressive and prosocial behaviour would be related at both the individual level and at the teacher/classroom level to teacher-student relationship quality. Finally, we hypothesised that between teacher/classroom level variation would be much lower when the same behaviours were assessed via self- rather than teacher-reports.

Method

Participants

The students in the current study were youth from the 3rd and 4th wave of the Zurich Project on Social Development from Childhood to Adulthood (z-proso). Z-proso is a longitudinal cohort study of the development of pro- and anti-social behaviour across childhood and adolescence. The study began in 2004 when the children were around 7 years old and beginning primary school. Children were invited to participate in the study if they attended one of 56 schools in Zurich, selected according to a stratified

random sampling procedure that took into account geographical location and size. The target sample was 1675 students, of which 1572 contributed data for at least one measurement wave. Analysis of non-participation and drop-out are reported in Eisner, Murray, Eisner & Ribeaud (2018) and Eisner & Ribeaud (2007). These studies suggested that youth with parents whose first language is not the official language of the study location (indicating immigrant status) are slightly under-represented in the study. However, the sample can otherwise be considered approximately representative of underlying same-aged population. Overall, data on 1001 students were available for the current analysis with some variation in sample size at the item level (see Table 1 in Results section for exact sample sizes). At the 3rd measurement wave (4th grade), the median age of the students was 10 and at the 4th measurement wave (5th grade), the median age was 11. Students were diverse in terms of ethnic and social background. The biggest proportion of primary caregivers of the students were born in Switzerland (53%), but many came from other nations including Serbia and Montenegro (6%), Sri Lanka (5%), Portugal (5%), Germany (5%), Italy (4%), Turkey (3%), Spain (2%), Macedonia (2%), Yugoslavia (2%), Bosnia-Herzegovina (2%), and 50 other nations accounting for 1% or less of the sample. In terms of socio-economic status, the mean ISEI household score (Ganzeboom, De Graaf & Treiman, 1992) for the students based on the n=921 for whom this data was available was 45.08 (SD=17.86). This score corresponds to occupational prestige levels associated with occupations such as routine clerical/sales work. The magnitude of SD of the ISES scores highlight the large variation in socioeconomic status within the sample.

At age 11, the students were rated by 250 teachers (37% male). At this stage, the teachers had been teaching the children for just less than two years. This is typical of the Zurich school system in which 3 grades of lower primary education with the same teacher (ages 7,8 ,9) are followed by 3 grades of middle level education with newly mixed classes and a new teacher (ages 10, 11, 12). The majority of children (>80%) retained the same teacher across grades 4 and 5. Teachers with at least seven participants in their class received a book voucher worth approximately 50USD as compensation for their participation. Further information on the sample, recruitment and assessment procedures for z-

proso can be found in previous publications (e.g. Eisner & Ribeaud, 2007; Eisner et al., 2018) and on the z-proso website: http://www.jacobscenter.uzh.ch/en/research/zproso/aboutus.html.

Measures

Aggression and *prosociality* were measured using items taken from the teacher and self-report forms of the Social Behavior Questionnaire (SBQ) (Tremblay et al., 1991). Aggression was measured using nine items referring to physical aggression (3 items), proactive aggression (3 items) and reactive aggression (3 items). Prosociality was measured using five items measuring behaviours indicative of empathy and helping. Both teacher and self-reported SBQ items were administered in German in a paper and pencil format and responses recorded on a five-point Likert scale from *never* to *very often*. Both were part of broader questionnaires assessing psychosocial functioning of the student. Abbreviated English translations are provided in results tables. Previous research has supported the reliability and validity of the measures (e.g. Lösel & Stemmler, 2012; Tremblay et al., 1991; Tremblay, Vitaro, Gagnon, Piché, & Royer, 1992). In the current sample, Murray, Eisner & Ribeaud. (2017) provided support for the factorial validity and reliability of the teacher-reported items using an item response theory approach. Murray, Obsuth, Eisner & Ribeaud, (2017) provided support for the reliability, factorial validity and metric invariance of the self-report items across adolescence. ω reliability values in the sample utilised in the current study are provided in Results tables for each dimension identified. All were .85 or above.

Teacher-student relationships were assessed using three items: 'My teacher treats me fairly', 'I get on well with my teacher' and 'My teacher helps me when necessary'. Both teacher-reported and student-reported teacher-student relationships data were available; however, we used the student-reported data to avoid inflated associations due to common rater bias. The measures were developed specifically for the z-proso study and were selected after piloting in a previous Swiss sample. Some of the items were drawn from a large German comparative study on youth violence (Wetzels, Enzmann, Mecklenburg & Pfeifer, 2001). Responses were provided on a four-point Likert scale from *false* to *true.* Items were administered as part of the same questionnaire as the SBQ self-reports, again in German

and in paper and pencil format. Reliability values for these items in the current sample were estimated from the ML-CFA models (described below) and were .76 or above.

Statistical procedure

Preliminary Analyses.

Item-level intraclass correlations (ICCs) and design effects were estimated for each item based on both teacher ratings and self-reports. The ICC expresses the ratio of between-teacher to total variance while the design effect expresses the extent of distortion of sampling error that would occur if assuming simple random sampling (e.g. Muthén & Satorra, 1995). Design effects increase with increasing ICCs and average cluster size. Typically, design effects less than 2 are considered small, with a single level model generally assumed adequate in these cases (e.g. Heck & Thomas, 2015).

We then obtained within-pooled and between-classroom correlation matrices for both the teacher and self-reported data by fitting multi-level models with saturated models at both the within and between level in *Mplus 7.13* (Muthén & Muthén, 2014; see Dyer, Hanges, & Hall, 2005). We conducted exploratory factor analyses of these matrices in order to guide the specification of the multi-level CFA models in subsequent steps. Parallel analysis, the minimum average partial test (MAP), and inspection of scree plots, together with the estimation and inspection of a range of factor solutions were used to guide the selection of the number of factors and their interpretation. EFAs were conducted using the *psych* package in R statistical software (R Core Team, 2016).

Multi-level CFAs.

We next fit multi- and single-level confirmatory factor analyses (CFAs) where specification of the between within and between structures were guided by the EFA results. Accordingly, if there was no evidence of a multi-level structure in any of the sets of items, multi-level CFAs were not fit. Scaling and identification were achieved by fixing latent variances to 1. ML-CFA models were estimated using robust maximum likelihood estimation in *Mplus 7.13* (MLR; Muthén & Muthén, 2014). Models were judged to fit well if CFI>.95, TLI>.95, RMSEA<.08 and SRMR <.05 (Hu & Bentler, 1999; Yu, 2002).

Level-specific fit statistics are also reported. Multi-level ω reliability values were computed to estimate the reliability of the within- and between-level constructs (see Geldhof, Preacher & Zyphur, 2014). The same identification constraints, estimation methods, fit criteria and reliability values are applicable to all subsequent ML-CFA and ML-structural equation model (ML-SEM) models (ML-SEM refers to models in which the ML-CFA is extended to include structural relations between latent factors and between latent factors and observed variables) described below.

Stability of between- and within-level ratings.

We assessed the stability of teacher-reported aggression and prosociality in the ML-CFAs. Stability was estimated by regressing the relevant latent factor at age 11 on the corresponding factor at age 10. The model specification is summarised in Figures 3 and 4, with the observed variables and between-level indicators omitted for clarity.

Correlates of Within- and Between- level ratings

We evaluated whether female teachers tend to evaluate children as less aggressive and more prosocial by regressing between-level aggression and prosocial factors at age 11 on teacher gender. We evaluated whether teachers who rated students as more aggressive or less prosocial tended to have poorer relationships with their students. Teacher-student relationship was also specified as a multi-level construct. At each both the within- and between- level it was defined as a single latent variable and its correlation with the same-level aggression and prosociality latent factor(s) estimated.

Results

Descriptive statistics

Descriptive statistics are provided in Table 1. The intraclass correlations (ICCs) for the 9 teacher-reported aggression items at age 11 ranged from .10 to .34, with all but one design effect >2. Similar ICCs and design effects were observed for these items in the age 10 teacher reports with design

effects here all >2. The same items as self-reports at age 11 yielded ICCs that were all <.01 with correspondingly small design effects (all <2).

ICCs and design effects for the prosociality items showed a similar pattern as those for the aggression items. At age 11 the teacher-reported ICCs ranged from .22 to .37 with design effects that were all >2. At age 10, the teacher-reported ICCs were slightly smaller but all design effects were >2. For the self-reported prosociality items at age 11, 2 ICCs were <.01 but the remaining were .05 or above, with a maximum ICC of .10. None of the design effects were >2.

ICCs for the teacher relationships items self-reported at age 11 ranged from .07 to .16. Only the item with the largest ICC had a design effect >2. The correlation matrix for all variables is provided as Supplementary Materials.

Multi-level EFAs

Exploratory factor analysis of the within-pooled correlation matrix for teacher-reported aggression at age 11 suggested that a 3-factor oblique model provided the best representation of the within-classroom level. These factors could be characterised as proactive aggression, reactive aggression and physical aggression. Analogous analyses of the between-level correlation matrix provided conflicting information on the optimal factor structure. Here parallel analysis suggested 1 factor to retain but MAP suggested 3. Inspection of a scree plot and loadings matrices for 1 to 3 factors under oblique and bi-factor rotations suggested the presence of a dominant general dimension with minor second and third factors, therefore, a single factor model was on balance preferred.

A one factor solution for prosociality was supported at the within- and between- levels, which was expected given the small number of items involved. For both aggression and prosociality, the same factor structures were replicated in the age 10 data. As the ICCs for the student-reported data were all low, we did not explore multi-level factor structures for these items.

Multi-level CFAs

The model specification for age 11 teacher-reported aggression is displayed in Figure 2 and for prosocality in Figure 3. These were based on the EFA results described above. In the aggression ML-CFA, it was also necessary to constrain the between-level residual variance of item 51 to a small positive value (0.01) to deal with Heywood cases. The aggression ML-CFA (one factor at the between-level, 3 factors at the within-level) yielded overall good fit by conventional criteria but poor fit at the between-level (TLI=.97; CFI=.96; RMSEA=.05; SRMR within-level=.04; SRMR between-level=.28). Model parameters are provided in Table 2. All loadings were salient (>.30 on a standardised scale) and the majority were >.70. The exceptions were the between-level general factor loading of items 33 (.40), 35 (.36) and 37 (.48). The within-level factors were highly correlated, with factor correlations between .60 and .72.

Modification indices and expected parameter changes were inspected to understand the poor fit at the between-level. These indicated a residual covariance between two proactive aggression items (items 51 and 52); however, fit barely improved with the addition of this parameter (between-level SRMR=.26). We suspect, therefore, given that there was some evidence of multidimensionality in the EFA, that a hierarchical model with both a general factor and group factors might be the overall best model to describe the between-classroom structure of aggression. However, we wanted to avoid capitalising on chance by making model modifications at the ML-CFA stage and thus retained our original model. Future studies with more comprehensive aggression assessments could explore both group and general factors of aggression at the between-classroom level.

The multi-level CFA for prosociality fit well overall and at both the between- and within- level (TLI=.95, CFI=.95, RMSEA=.05, SRMR within-level=.03; SRMR between-level=.02). Model parameters are provided in Table 3. All standardised loadings at both the between- and within level were >.60.

As with the EFAs, given that the ICCs for the student-reported data were all low, we did not fit ML-CFAs to the student-reported items.

Stability of between- and within-level aggression and prosociality.

We confirmed that the same multi-level structure could be fit to the teacher-reported aggression items at the previous measurement wave (age 10). Parameter estimates are provided in Table 2. Within- and between-level factor loadings were almost all salient (>.30 on a standardised scale). In fact, the majority were >.70. The exceptions were the loadings of item 33,37,37 and 53 on the general between-level factor. These were .35, .33, .61 and .17 respectively.

We then tested the stability of the between- and within-teacher/classroom factors in these models. We here restricted our analyses to the children who had the same teacher at both time points. This was 829 children rated by 161 teachers. We began by testing configural invariance. Reactive, proactive and physical aggression factors were allowed to correlate at the within-level levels. Similarly, residual covariances between the same item measured at age 10 and 11were freely estimated. Scaling and identification were achieved by fixing the latent variances of the age 10 factors to 1 and fixing one loading per factor equal across age 10 and 11. It was also necessary to constrain the between-level residual variances of SBQ item 34 at both age 10 and 11 to a small positive value to overcome Heywood cases. The configural model fit well (CFI=.94, TLI=.92, RMSEA=.050). The addition of metric invariance constraints improved model fit (CFI=.95, TLI=.93, RMSEA=.047). Scalar or residual invariance was not required as latent means were not compared and all comparisons were done within a latent model (as opposed to using summed or average scores; e.g. see Liu et al., 2016). Using the metric invariant model, the standardised autoregressive parameters for within-level proactive, reactive and physical aggression factors were respectively .56, .54 and .60. Between-level stability of general aggression was .38.

An analogous set of analyses were conducted for prosociality after confirming that the factor structure supported at age 11 was also supported at age 10. Table 3 provides the standardised parameter estimates and shows that all loadings were >.60. The fit of the configural model was acceptable (CFI=.91, TLI=.85, RMSEA=.076). The addition of metric invariance constraints improved model fit (CFI=.92, TLI=.88, RMSEA=.069). Based on the model with metric invariance constraints, within-level stability of prosociality was .59. The between-level stability was .55.

Relations of between- and within-level factors with external criteria

Regressing between-level aggression at age 11 on the gender of the rater suggested that male teachers were no less likely to attribute aggressive behaviours to students than female teachers (β =0.13; *p*=.64). There was also no significant effect of gender on between-level prosociality (β= -0.06; *p*=.48).

In the model with teacher-student relationships it was necessary to constrain the between-level residual of the items measuring teacher fairness and getting along with the teacher to small positive values to deal with Heywood cases. Standardised factor loadings for the three teacher relationship items were all >.50 and statistically significant (*p*<.001) at the within-level, with ω =.76. At the between-level all standardised loadings were >.90 and significant (*p*<.001), with ω =.99. There was also no significant between-level correlation between aggression and teacher-student relationships (*r*=-.20; *p*=.26). There were, however, significant correlations between teacher-student relationships and proactive (*r*=-.36, *p*<.001), reactive (*r*=-.27, *p*<.001) and physical aggression (*r*=-.32, *p*<.001) at the within-teacher/classroom level. Similar to aggression, there was also no significant association between teacher-student relationships and prosociality at the between-teacher/classroom level (*r*=.13, *p*=.50) but a significant association at the within-teacher/classroom level (*r*=.25, *p*<.001).

Discussion

In this study, we sought to explore the nature and source of variation in student behaviour ratings occurring at the between-teacher/classroom level. We found evidence for strong between-factor structures for teacher-rated prosocial and aggressive behaviours. In terms of the dimensions along which teacher/classroom-level differences in ratings varied, both aggression and prosociality between-teacher/classroom structures were best characterised by a single dimension, although some potential multidimensionality in the former was evident. For aggression, reactive, proactive and physical aggression dimensions could be distinguished at the within-teacher/classroom level. For prosociality, the within-teacher/classroom structure was best characterised as unidimensional.

Within- and between- teacher factor stability

Both the within- and between-teacher constructs showed stability over time. The within-teacher/classroom stabilities of proactive, reactive, and physical aggression (.54-.60) and prosociality (.59) were in line with those of past studies and suggested that individual differences in both traits shows moderate to strong stability over the course of development (e.g. Eisenberg et al., 2002; Tuvblad, Raine, Zheng & Baker, 2009; Verhulst & van der Ende, 1995). No studies – to the best of our knowledge – have previously examined the stability of these constructs at the between-classroom/teacher level. No benchmarks for interpreting between-teacher/classroom stability currently exist but we see no strong reason to assume that constructs at the between-teacher/classroom level would be inherently more or less stable than those at the within-teacher/classroom level when the class composition remains the same over the period studied. In terms of the specific constructs studied here, however, the between-teacher/classroom stability of aggression (.38) was lower than its within-teacher/classroom stabilities. This suggests a trait-like component but in the context of substantial reshuffling in classroom-level aggressive behaviour. The reason for the instability is not clear. The between-teacher/classroom stability of prosociality (.55) was, however, close in magnitude to its within-teacher/classroom stability suggesting that differences between classrooms in prosocial behaviour (or teacher perceptions thereof) exhibit a substantial trait-like component. Identifying the influences that serve to maintain high or low classroom-level reported prosociality and aggression represents an interesting future direction. Similarly, extending the simple autoregressive models presented here to identify predictors of teacher/classroom-level change over time (within or across school years) is likely to be valuable for generating evidence that can be incorporated into teacher/classroom-level interventions (see Marsh et al., 2012 for a more general discussion of the promise of ML-CFA in school-based research).

Correlates of between-classroom levels of aggression and prosociality

We examined two correlates of between-classroom levels of aggression and prosociality: teacher gender and teacher-student relationships (e.g. Spilt et al., 2012). Contrary to our hypothesis, female teachers did not provide more overall favourable ratings of their students than male teachers either on aggression or prosociality. Previous studies have yielded mixed results on gender differences in teacher ratings of students. Some studies have found that female teachers provide more favourable

ratings (e.g. Spilt et al., 2012), others have found no difference (e.g., Krkovic, Greiff, Kupiainen, Vainikainen & Hautamäki, 2014). Others still have found that female teachers provide more favourable of only some traits such as prosociality and emotion regulation, but not others such as concentration problems, internalising and disruptive behaviour problems (Pas & Bradshaw, 2014). We speculate that teacher gender effects may vary by culture and depend on factors such as overall male:female teacher ratios; however, future research will be required to determine the source of inconsistencies across studies.

Prosociality and aggression were also not significantly associated with teacher relationships at the between-teacher/classroom level, although both were associated with teacher relationships at the within-teacher/classroom level. The within-teacher/classroom association replicates a growing body of research showing links between teacher-student relationships and student behaviour; an association that appears to reflect bidirectional effects of the relationship on behaviour and behaviour on relationship (e.g., Theimann, 2016). However, only a handful of studies have examined teacher-student relationships from a multi-level perspective. Our study is broadly consistent with these in identifying systematic variation in teacher-student relationships at both the student and teacher/classroom level (e.g. Spilt et al., 2012). Our results suggest that the former level is more important as regards prosocial and aggressive student behaviour. They do not, however, imply that student-derived factors are necessarily more important than teacher-derived factors because variance due to the interactive effect of teacher and student (e.g. teacher responses that are unique to a given student because of their particular pattern of behaviour) will be captured at within-classroom/teacher level as well. To the extent that these interactive effects are important, teacher characteristics will also be. We propose that disaggregation of these two sources of variance could be achieved in studies that examine relationships between multiple students and multiple teachers either over school years or in the latter stages of schooling where students routinely interact with more than one teacher.

Student self-reports

Finally, we attempted to replicate the above-described analyses with student self-reports; however, self-reports did not show evidence of substantial between-classroom structure for either prosociality or aggression (ICCs <.05; Dyer et al., 2005). ML-CFA was thus not possible with these measures. Discrepancies in magnitudes of level-2 variance between teacher versus student reports have been observed in previous studies. In a study of school exclusion, for example, Obsuth, Sutherland et al. (2016) reported that although there was substantial clustering by school in teacher reports, there was little evidence of clustering in behaviour ratings by school in the corresponding youth self-reports. There are a number of possible interpretations for the discrepancy we observed between teacher and student ICCs. One possibility is that the larger amount of between teacher/classroom variation in teacher reports reflects the effect of response style variations across teachers (e.g., Podsakoff, MacKenzie & Podsakoff, 2012). The other main possibility is that it reflects the fact that while teacher ratings primarily reflect school, especially classroom-based behaviours (e.g. De Los Reyes, Henry, Tolan & Wakschlag, 2009); students are likely to draw on their behaviour across multiple contexts. If shared classroom environments create similar classroom-specific behaviours that do not generalise to other contexts, teacher and student ratings would be likely to differ in the amount of clustering observed. This interpretation would be consistent with the fact that when rating a classroom-specific construct i.e. teacher relationships, student ratings showed a much higher level of level-2 variance than when rating their individual aggressive and prosocial behaviour. However, there are also factors that could artificially attenuate between-teacher/classroom variance in self-reports. For example, youth may exaggerate the difference between themselves and their classmates because these peers are predominant in the reference frame by which they judge their own behaviour. Self-reports of aggression may also have large amounts of measurement error when collected at younger ages (e.g. Murray, Obsuth et al., 2017). Independent observations of classroom behaviours could help to disentangle these various possibilities. We would anticipate that all of the above explanations contribute to some extent.

Implications

If a large component of level-2 variance reflects teacher rating biases, it would be important to consider whether current survey practices permit teachers to rely too heavily on their implicit theories

(e.g. Abikoff et al., 1993) or other response biases rather than their observations when providing data about student behaviour. This could potentially be remedied through rating instructions to teachers that raise awareness of potential implicit biases, and which encourage them to attend only to their direct observations of behaviour. Further, the extent to which teacher perceptions of student behaviour reflect characteristics of the teacher rather than the true behaviour of students has important implications for supporting student academic and psychosocial development. It has been suggested that the extent to which youth feel they are viewed positively by their teacher influences their future success and well-being (e.g. Kellam & Rebok, 1992; Krkovic et al., 2014). The hypothesis holds that when teachers hold unfavourable perceptions of children, this manifests in their interactions, impacting on the extent to which their development is positively supported. Dobbs and Arnold (2009), for example, found that teacher interactions with students depended on their perceptions of the student behaviour over and above the student's actual behaviour. To the extent that negative teacher perceptions reflect implicit biases rather than actual student behaviours, teacher perceptions may represent a key target for intervention to optimise student development.

Limitations

Finally, it is important to consider the limitations of the current study. A primary limitation is the brevity of the measures of prosociality and aggression available. While the CFA analyses suggested that these measures showed high reliability, more comprehensive measures would have provided opportunities for identifying within- and between-rater structures pertaining to a broader range of indicators of aggressive and prosocial behaviours. It is possible that a more diverse set of indicators would have revealed multi-factor between-teacher/classroom structures. Similarly, we examined only two correlates of between- and/or within-classroom aggression and prosociality (i.e. teacher gender and teacher-pupil relationships). A range of other potential predictors and outcomes would be of interest in relation to these constructs, including, for example, the effects of interventions, teacher stress and mental health, and teacher attitudes. In addition, only a subsample of each teacher's class was measured, thus we could not estimate full classroom context. Future studies that have measurements on entire classrooms of students would be in a better position to do this. Finally, we did not have enough

level-2 units to split the sample for the EFA and CFA analyses. Thus, we could not validate our EFA results solution in an independent sample. Future replications in independent samples will thus be valuable for assessing the generalisability of the factor solutions developed in the current study.

Conclusion

There is systematic variance in teacher ratings of prosocial and aggressive behaviour both between and within teachers/classrooms. Studies aiming to identify the source of variation between teachers/classrooms would represent a fruitful line of research to inform teacher/classroom- level intervention targets to improve classroom-wide student behaviour. We examined one potential factor and though related at the within-teacher/classroom level, teacher relationships did not significantly correlate with between-teacher/classroom prosociality and aggression.

References

Abikoff, H., Courtney, M., Pelham Jr., W. E., & Koplewicz, H. S. (1993). Teachers' ratings of disruptive behaviors: The influence of halo effects. *Journal of Abnormal Child Psychology*, *21*, 519-533.

Aloe, A. M., Shisler, S. M., Norris, B. D., Nickerson, A. B., & Rinker, T. W. (2014). A multivariate meta-analysis of student misbehavior and teacher burnout. *Educational Research Review, 12,* 30-44.

Arens, A. K., & Morin, A. J. (2016). Relations between teachers' emotional exhaustion and students' educational outcomes. *Journal of Educational Psychology, 108*, 800-813.

Barth, J. M., Dunlap, S. T., Dane, H., Lochman, J. E., & Wells, K. C. (2004). Classroom environment influences on aggression, peer relations, and academic focus. *Journal of School Psychology*, *42*, 115-133.

Berg-Nielsen, T. S., Vika, A., & Dahl, A. A. (2003). When adolescents disagree with their mothers: CBCL-YSR discrepancies related to maternal depression and adolescent self-esteem. *Child: care, health and development*, *29*, 207-213.

Downer, J. T., Stuhlman, M., Schweig, J., Martínez, J. F., & Ruzek, E. (2015). Measuring effective teacher-student interactions from a student perspective: A multi-level analysis. *The Journal of Early Adolescence*, *35*, 722-758.

Dunn, E. C., Masyn, K. E., Jones, S. M., Subramanian, S. V., & Koenen, K. C. (2015). Measuring psychosocial environments using individual responses: an application of multilevel factor analysis to examining students in schools. *Prevention Science, 16*, 718-733.

De Los Reyes, A., Henry, D. B., Tolan, P. H., & Wakschlag, L. S. (2009). Linking informant discrepancies to observed variations in young children's disruptive behavior. *Journal of Abnormal Child Psychology,* 37, 637-652.

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly*, *16*, 149-167.

Eisner, N., Murray, A.L., Eisner, M., Ribeaud. (2018). An Analysis of Non-response and Attrition in the Zurich Project on Social Development from Childhood to Adulthood (z-proso). *International Journal of Behavioral Development*. In press.

Eisner, M., & Ribeaud, D. (2007). Conducting a criminological survey in a culturally diverse context lessons from the Zurich project on the social development of children. *European Journal of Criminology*, *4*, 271-298.

Eisenberg, N., Guthrie, I. K., Cumberland, A., Murphy, B. C., Shepard, S. A., Zhou, Q., & Carlo, G. (2002). Prosocial development in early adulthood: a longitudinal study. *Journal of Personality and Social Psychology*, *82*, 993-1006.

Fraser, M. W., Galinsky, M. J., Smokowski, P. R., Day, S. H., Terzian, M. A., Rose, R. A., & Guo, S. (2005). Social information-processing skills training to promote social competence and prevent aggressive behavior in the third grades. *Journal of Consulting and Clinical Psychology*, *73*, 1045.

Ganzeboom, H. B., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21,* 1-56.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19,* 72-91.

Hartman, C. A., Rhee, S. H., Willcutt, E. G., & Pennington, B. F. (2007). Modeling rater disagreement for ADHD: are parents or teachers biased?. *Journal of Abnormal Child Psychology*, *35*, 536-542.

Harty, S. C., Miller, C. J., Newcorn, J. H., & Halperin, J. M. (2009). Adolescents with childhood ADHD and comorbid disruptive behavior disorders: aggression, anger, and hostility. *Child Psychiatry and Human Development, 40,* 85-97.

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus*. Routledge.

Hopf, D., & Hatzichristou, C. (1999). Teacher gender-related influences in Greek schools. *British Journal of Educational Psychology*, *69*, 1-18.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1-55.

Kellam, S. G., & Rebok, G. W. (1992). Building developmental and etiological theory through epidemiologically based preventive intervention trials. In J. McCord, & R. E. Tremblay (Eds.), *Preventing antisocial behavior: Interventions from birth through adolescence* (pp. 162–195). New York: Guilford Press.

Krkovic, K., Greiff, S., Kupiainen, S., Vainikainen, M. P., & Hautamäki, J. (2014). Teacher evaluation of student ability: What roles do teacher gender, student gender, and their interaction play? *Educational Research*, *56*, 244-257.

Laninga-Wijnen, L., Harakeh, Z., Dijkstra, J. K., Veenstra, R., & Vollebergh, W. (2017). Aggressive and prosocial peer norms: Change, stability, and associations with adolescent aggressive and prosocial behavior development. *The Journal of Early Adolescence*. Online First.

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods, 22*, 486-506.

Lösel, F., & Stemmler, M. (2012). Preventing child behavior problems in the Erlangen-Nuremberg Development and Prevention Study: Results from preschool to secondary school age. *International Journal of Conflict and Violence*, *6*, 214-224.

Malti, T., Ribeaud, D., & Eisner, M. P. (2011). The effectiveness of two universal preventive interventions in reducing children's externalizing behavior: A cluster randomized controlled trial. *Journal of Clinical Child & Adolescent Psychology*, *40*, 677-692.

Menesini, E., Palladino, B. E., & Nocentini, A. (2015). Emotions of moral disengagement, class norms, and bullying in adolescence: A multilevel approach. *Merrill-Palmer Quarterly, 61,* 124-143.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*, 106-124.

Morin, A. J., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *The Journal of Experimental Education*, *82*, 143-167.

Murray, A. L., Eisner, M., Ribeaud, D. (2017). In search of trans-diagnostic

dimensional measures of childhood and adolescent psychopathology: An analysis of the Social Behavior Questionnaire. *European Journal of Psychological Assessment.* In Press.

Murray, A. L., Obsuth, I., Eisner, M., & Ribeaud, D. (2017). Evaluating Longitudinal Invariance in Dimensions of Mental Health Across Adolescence: An Analysis of the Social Behavior Questionnaire. *Assessment*. Online First.

Murray, A. L., Obsuth, I., Eisner, M., & Ribeaud, D. (2016). Shaping aggressive personality in adolescence: Exploring cross-lagged relations between aggressive thoughts, aggressive behaviour and self-control. *Personality and Individual Differences, 97*, 1-7.

Murray, A. L., Obsuth, I., Zirk-Sadowski, J., Ribeaud, D., & Eisner, M. (2016). Developmental relations between ADHD symptoms and reactive versus proactive aggression across childhood and adolescence. *Journal of Attention Disorders.* Online First.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376-398.

Muthén, L. K., & Muthén, B. O. (2014). *Mplus User's Guide*. (7th ed.). Muthén & Muthén, Los Angeles.

O'Brennan, L. M., Bradshaw, C. P., & Furlong, M. J. (2014). Influence of classroom and school climate on teacher perceptions of student problem behavior. *School Mental Health*, *6*, 125-136.

Obsuth, I., Murray, A. L., Malti, T., Sulger, P., Ribeaud, D., & Eisner, M. (2017). Non-bipartite propensity score analysis of the effects of teacher-student relationships on adolescent problem and prosocial behavior. *Journal of Youth and Adolescence,* 1-27.

Obsuth, I., Sutherland, A., Cope, A., Pilbeam, L., Murray, A. L., & Eisner, M. (2016). London Education and Inclusion Project (LEIP): Results from a Cluster-Randomized Controlled Trial of an Intervention to Reduce School Exclusion and Antisocial Behavior. *Journal of Youth and Adolescence*, 1-20.

Pakarinen, E., Silinskas, G., Hamre, B. K., Metsäpelto, R. L., Lerkkanen, M. K., Poikkeus, A. M., & Nurmi, J. E. (2017). Cross-lagged associations between problem behaviors and teacher-student relationships in early adolescence. *The Journal of Early Adolescence*. Online First.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539-569.

Pas, E. T., & Bradshaw, C. P. (2014). What affects teacher ratings of student behaviors? The potential influence of teachers' perceptions of the school environment and experiences. *Prevention Science*, *15*, 940-950.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Salmivalli, C., Voeten, M., & Poskiparta, E. (2011). Bystanders matter: Associations between reinforcing, defending, and the frequency of bullying behavior in classrooms. *Journal of Clinical Child & Adolescent Psychology*, *40*, 668-676.

Spilt, J. L., Koomen, H. M., & Jak, S. (2012). Are boys better off with male and girls with female teachers? A multilevel investigation of measurement invariance and gender match in teacher–student relationship quality. *Journal of School Psychology*, *50*, 363-378.

Theimann, M. (2016). School as a space of socialization and prevention. *European Journal of Criminology, 13,* 67-91.

Thomas, D. E., & Bierman, K. L. (2006). The impact of classroom aggression on the development of aggressive behavior problems in children. *Development and Psychopathology*, *18*, 471-487.

Thomas, D. E., Bierman, K. L., & Powers, C. J. (2011). The influence of classroom aggression and classroom climate on aggressive–disruptive behavior. *Child Development*, *82*, 751-757.

Tremblay, R. E., Loeber, R., Gagnon, C., Charlebois, P., Larivee, S., & LeBlanc, M. (1991). Disruptive boys with stable and unstable high fighting behavior patterns during junior elementary school. *Journal of Abnormal Child Psychology*, *19*, 285-300.

Tremblay, R. E., Vitaro, F., Gagnon, C., Piché, C., & Royer, N. (1992). A prosocial scale for the Preschool Behaviour Questionnaire: Concurrent and predictive correlates. *International Journal of Behavioral Development*, *15*, 227-245.

Tuvblad, C., Raine, A., Zheng, M., & Baker, L. A. (2009). Genetic and environmental stability differs in reactive and proactive aggression. *Aggressive Behavior*, *35*, 437-452.

Wetzels, P., Enzmann, D., Mecklenburg, E., & Pfeiffer, C. (2001). Jugend und Gewalt: Eine repräsentative Dunkelfeldanalyse in München und acht anderen deutschen Städten. Interdisziplinäre Beiträge zur kriminologischen Forschung: Vol. 17. Baden-Baden: Nomos.

Verhulst, F. C., & Van Der Ende, J. (1995). The eight-year stability of problem behavior in an epidemiologic sample. *Pediatric Research, 38,* 612-617.

Youngstrom, E., Loeber, R., & Stouthamer-Loeber, M. (2000). Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting and Clinical Psychology*, *68*, 1038.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Vol. 30). Los Angeles: University of California, Los Angeles.

Tables

Table 1:

Descriptive Statistics, ICCs and design effects for all variables

		Age 10 Teacher Report					Age 11 Teacher Report					Age 11 Self-report				
Item	**Content**	**n**	**Mean**	**SD**	**ICC**	**Design effect**	**n**	**Mean**	**SD**	**ICC**	**Design effect**	**n**	**Mean**	**SD**	**ICC**	**Design effect**
							Aggression									
33	Gets into fights	999	1.46	0.83	0.22	2.77	999	1.37	0.74	0.10	1.97	999	1.43	0.80	<.01	0.92*
34	Physically attacks	999	1.4	0.77	0.17	2.63	1000	1.34	0.71	0.13	2.19	994	1.43	0.75	<.01	0.91*
35	Kicks, bites, hits	1001	1.32	0.71	0.24	2.98	999	1.25	0.63	0.16	2.47	1000	1.52	0.79	<.01	0.64*
37	Threatens	1000	1.29	0.66	0.16	2.48	999	1.27	0.64	0.18	2.49	994	1.12	0.45	<.01	0.87*
51	Tries to dominate	995	1.57	0.91	0.20	2.55	996	1.54	0.88	0.13	2.21	997	1.35	0.68	<.01	1.05
52	Scares other children	994	1.32	0.71	0.23	2.75	997	1.3	0.71	0.23	2.90	999	1.21	0.56	<.01	1.90
53	Aggressive if teased	997	1.95	1.12	0.28	3.10	998	1.89	1.07	0.28	3.02	994	2.73	1.09	<.01	1.30
54	Aggressive if something taken	996	1.94	1.1	0.31	3.44	995	1.86	1.01	0.34	3.50	991	1.38	0.69	<.01	0.84*
55	Aggressive if contradicted	997	1.68	0.95	0.29	3.23	997	1.64	0.91	0.31	3.28	1001	1.71	0.84		1.13
							Prosociality									
41	Volunteers to help	995	2.93	1.10	0.15	2.05	988	3.00	1.17	0.22	2.51	1001	3.34	1.09	0.10	1.87

42	Tries to stop disputes	988	3.02	1.04	0.25	2.85	991	3.07	1.10	0.28	3.08	993	3.48	1.17	0.06	1.59
43	Tries to help someone who is hurt	978	3.41	0.98	0.26	2.99	980	3.46	1.06	0.37	3.83	998	4.13	0.96	0.05	1.32
46	Comforts upset peer	980	3.30	0.96	0.23	2.78	985	3.37	1.00	0.30	3.23	999	3.98	1.01	<.01	1.60
49	Shares things	977	3.38	0.84	0.23	2.75	991	3.43	0.89	0.26	2.94	1000	3.89	0.98	<.01	1.17
	Teacher relationships															
1	Treats me fairly	-	-	-	-		-	-	-	-	2.08	992	2.44	0.74	0.16	2.08
2	Get on well	-	-	-	-		-	-	-	-	1.82	993	2.54	0.68	0.13	1.82
3	Teacher helps me	-	-	-	-		-	-	-	-	1.49	994	2.50	0.70	0.07	1.49

***Note*. ***ICCs where estimated as small negative values.

Table 2: Standardised ML-CFA factor loading and factor correlation estimates for aggression

		Age 10 teacher report				Age 11 teacher report			
		Factor Loadings							
		Within			**Between**	**Within**			**Between**
Item	**Item content**	**PA**	**PhyA**	**RA**	**GA**	**PA**	**PhyA**	**RA**	**GA**
33	Gets into fights		.91		.35		.90		.40
34	Physically attacks		.97		.96		.95		.85
35	Kicks, bites, hits		.90		.33		.86		.36
37	Threatens	.83			.61	.84			.48
51	Tries to dominate	.77			.82	.74			.94
52	Scares other children	.87			.77	.88			.75
53	Aggressive if teased			.91	.17			.90	.94
54	Aggressive if something taken			.89	.88			.90	.87
55	Aggressive if contradicted			.92	.94			.86	.87
	***ω* reliability**	.85	.95	.93	.94	.85	.93	.92	.95
		Factor correlations							
	PA	1	-	-	-	1	-	-	-
	RA	.71	1	-	-	.71	1	-	-
	PhyA	.76	.70	1	-	.72	.69	1	-

Note. RA= reactive aggression; PA=proactive aggression; PhyA= physical aggression; GA=general aggression.

Table 3: Standardised ML-CFA factor loadings for prosociality

		Teacher reports age 10		**Teacher reports age 11**	
Item	**Item content**	**Within**	**Between**	**Within**	**Between**
41	Volunteers to help	.70	.82	.73	.91
42	Tries to stop disputes	.80	.98	.84	.94
43	Tries to help someone who is hurt	.85	.93	.87	.88
46	Comforts upset peer	.79	.84	.80	.86
49	Shares things	.65	.73	.63	.82
ω					
reliability		.87	.94	.88	.95

Figures

Figure 1:

Example ML-CFA model for aggression

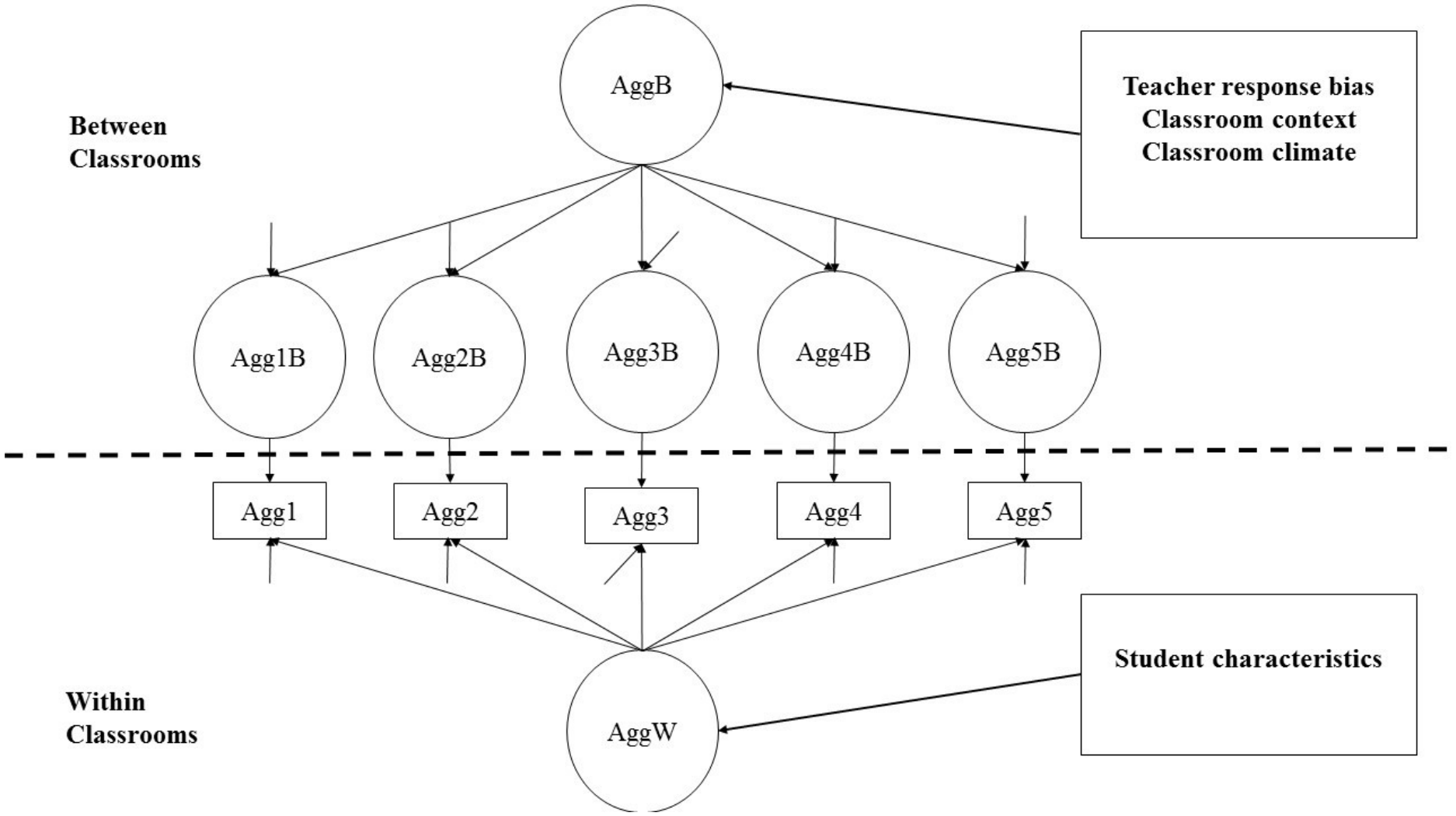
AggB
Teacher response bias
Classroom context
Classroom climate
Between
Classrooms
Agg1B
Agg2B
Agg3B
Agg4B
Agg5B
Agg1
Agg2
Agg3
Agg4
Agg5
Student characteristics
Within
Classrooms
AggW

Figure 2: ML-CFA Model Specification for Aggression

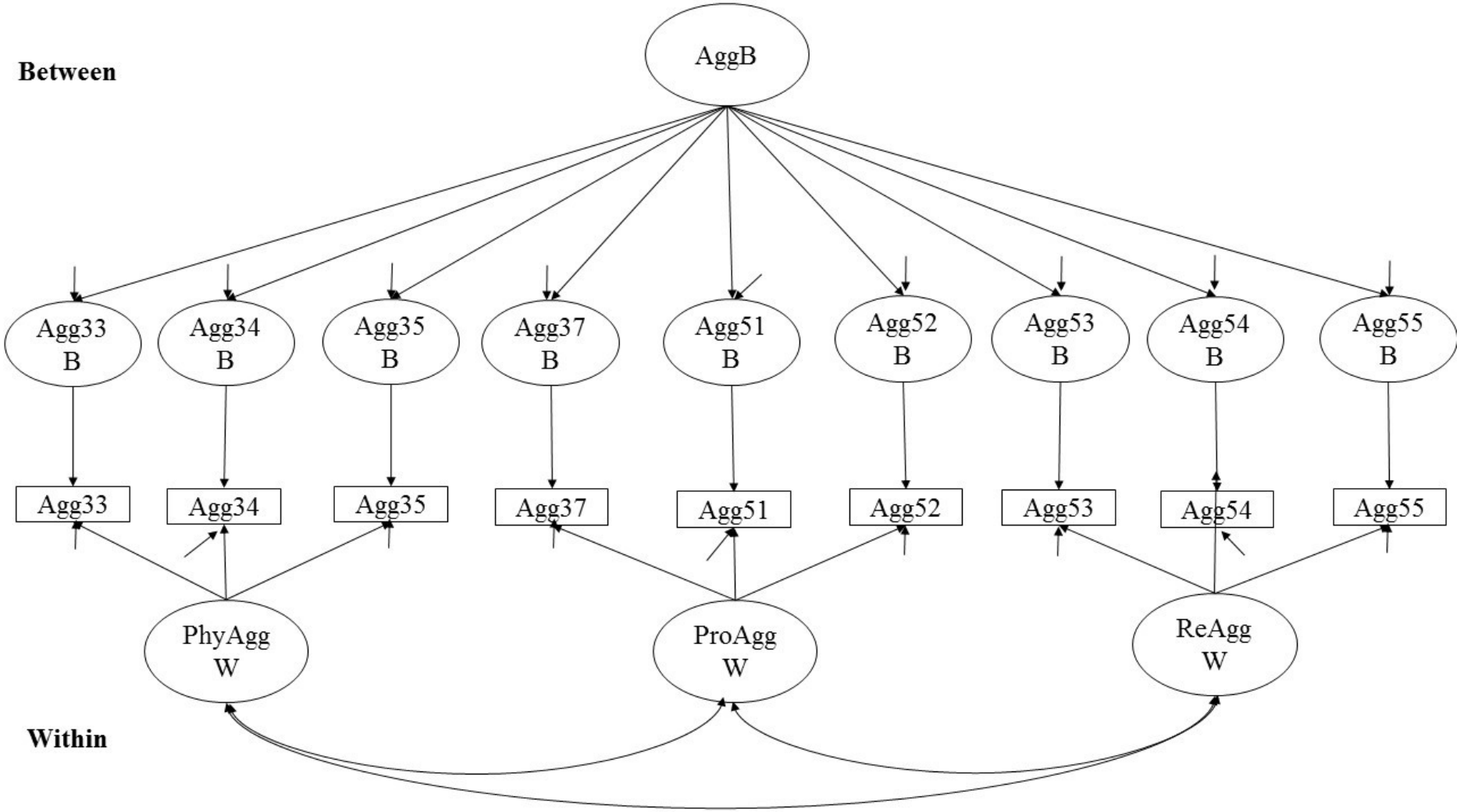

Figure 3: ML-CFA for Model Specification for Prosociality

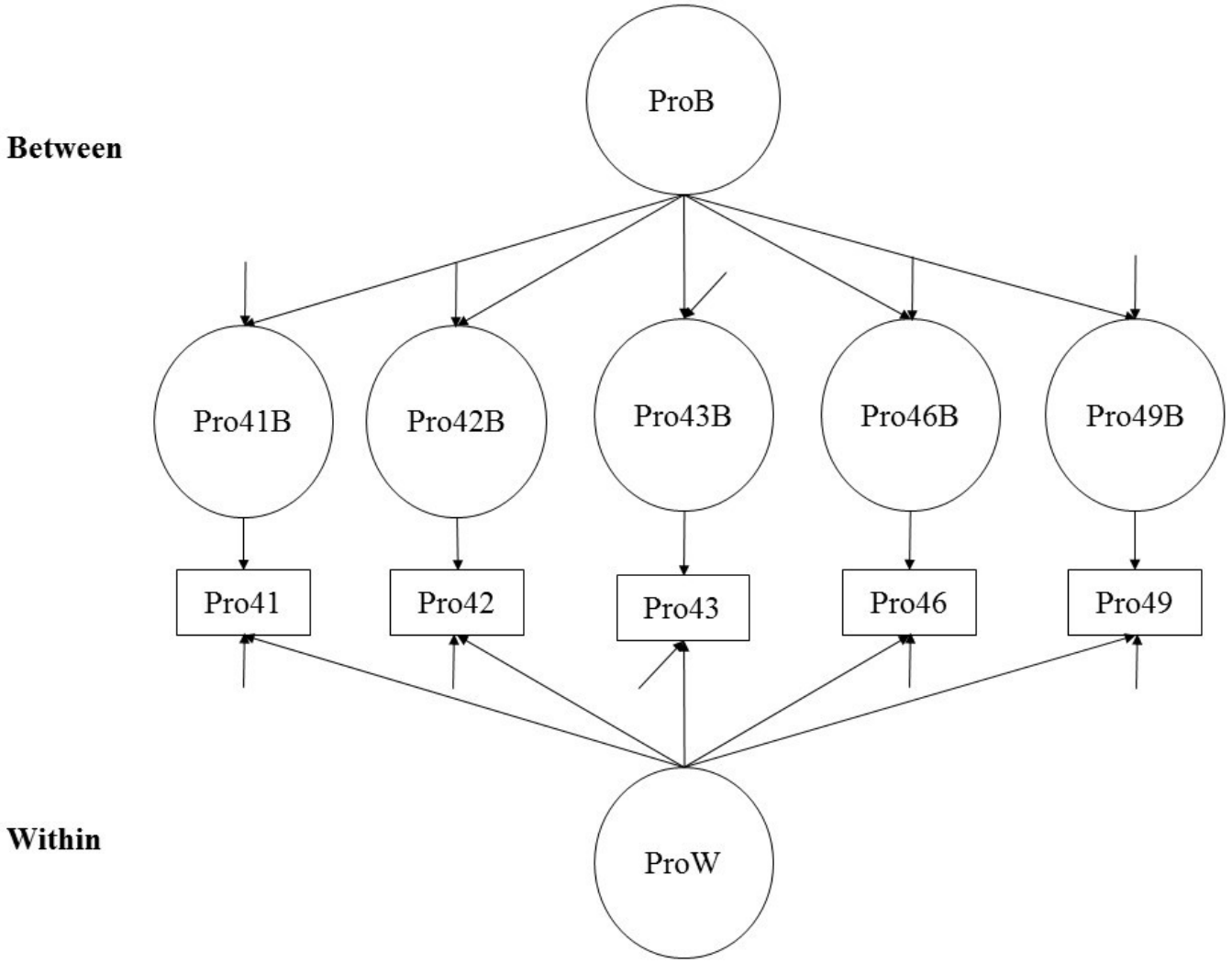

Figure 4: ML-CFA estimating the stability of Aggression

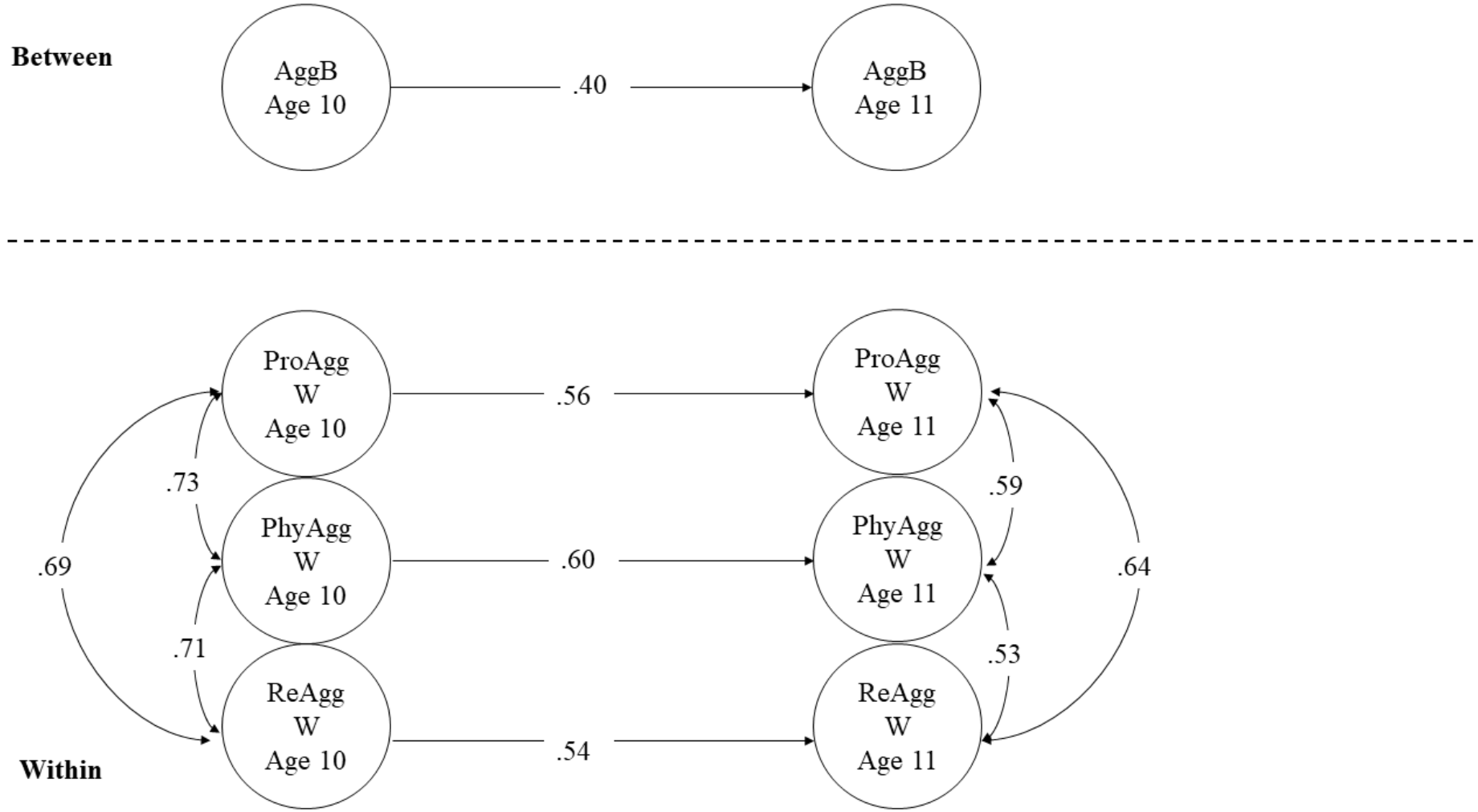

Figure 5: ML-CFA estimating the stability of Prosociality